



Classification

Ezequiel López-Rubio and Enrique Domínguez
Department of Computer Languages and
Computer Science
University of Málaga, Spain



Contents

- 1. Introduction
- 2. Fundamentals
- 3. Performance evaluation
- 4. Naïve Bayes classifiers
- 5. Conclusion



1. INTRODUCTION

1. Introduction

Motivation

- ❑ Many problems in health sciences are classification tasks:
 - Medical diagnosis
 - Treatment selection
 - Relapse prediction
 - Survival prediction





1. Introduction

Overview

- ❑ In this unit we study methods to solve **classification problems**
- ❑ Special attention is devoted to **performance evaluation**, since it must be done carefully in order to avoid overly optimistic results
- ❑ A simple but powerful kind of classifiers is considered in detail, namely the **naïve Bayes classifiers**



2. FUNDAMENTALS

2. Fundamentals

Supervised learning

- Given a query point \mathbf{x}_q in an input space of dimension D , the task of supervised learning is to **discover** an estimator $y_q = h(\mathbf{x}_q)$ of an unknown function f from a set of **examples** $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$ of size N . Let V be the range of f :
 - If V is a continuous set (for example, an interval of real numbers), then we have a **regression** problem
 - If V is a small set of discrete values, then we have a **classification** problem. We only study these problems here

2. Fundamentals

Classes and hypotheses

- Each element of V is named a **class**
 - If the number of classes is $C=2$, we say that it is a **binary classification** problem. Otherwise, it is **multiclass**
 - We must choose an estimator function h (also called hypothesis) from a **hypothesis space**
 - An estimator performs well if it **generalizes** well, i.e. if it correctly predicts the value of y for novel examples
 - It is also desired that the estimators are as simple as possible. This principle is called **Occam's razor**

2. Fundamentals

Probabilistic classifiers

- Often we want to design a classifier that outputs the **probabilities** that a given sample \mathbf{x}_q belongs to each of the classes y , $P(y \mid \mathbf{x}_q)$
 - This kind of classifiers are called probabilistic
- In these cases, the estimated class $y_q \in V$ is defined as the **most likely** one:

$$y_q = \arg \max_{y \in V} P(y \mid \mathbf{x}_q)$$



2. Fundamentals

Splitting the example set

- The available set of examples must be split into three sets:
 - The **training set**, which is supplied to one or more learning algorithms to obtain estimators h_i
 - The **validation set**, which is used to choose the best estimator among these
 - The **test set**, which is used to evaluate the performance of the chosen estimator



3. PERFORMANCE EVALUATION



3. Performance evaluation

General performance measures

- The **accuracy** is the number of correctly predicted examples divided by the number of examples
- The **Rand index** is the number of pairs of examples which are correctly predicted to belong to the same class, plus the number of pairs of examples which are correctly predicted to belong to different classes, divided by the number of pairs of examples
- Both measures lie in the interval $[0,1]$, and higher is better (1 means perfect classification)

3. Performance evaluation

Confusion matrix

- The **confusion matrix** has size $C \times C$, and its (i,j) element is the number n_{ij} of examples of class j which are predicted to belong to class i
 - The performance for the j -th class can be assessed by studying the j -th column of the matrix

	Actual class			
		1	...	C
	1	n_{11}	...	n_{1C}

	C	n_{C1}	...	n_{CC}
Predicted class				

3. Performance evaluation

Binary classification measures

- For a binary classification problem where class 1 is called the **positive class**, and class 2 is called the **negative class**, we define:
 - True positives (TP)= n_{11}
 - True negatives (TN)= n_{22}
 - False positives (FP , also called Type I errors)= n_{12}
 - False negatives (FN , also called Type II errors)= n_{21}
- Note that:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

3. Performance evaluation

Binary classification measures

- From the preceding we define: **precision** (also called positive predictive value, higher is better), **fallout** (or false positive rate, lower is better), **recall** (also called sensitivity or true positive rate, higher is better), and **F-measure** (higher is better)

$$Precision = \frac{TP}{TP + FP}$$

$$Fallout = \frac{FP}{FP + TN}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F - measure = 2 \frac{Precision \cdot Recall}{Precision + Recall}$$



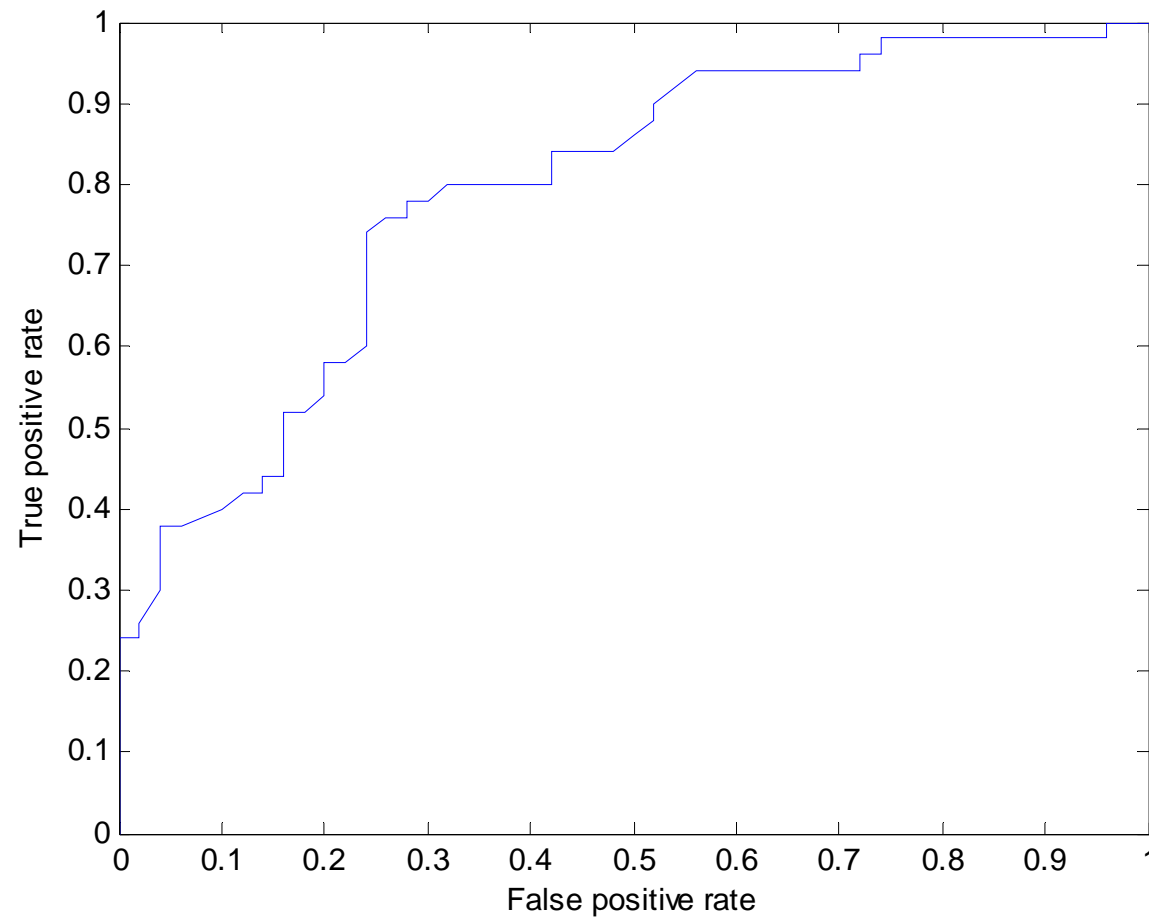
3. Performance evaluation

Receiver operating characteristic (ROC)

- ❑ For binary classification only. It is the plot of the true positive rate versus the false positive rate, as some discrimination parameter is varied
- ❑ The higher the **area under the ROC curve**, the better; it is a number in the interval $[0,1]$
- ❑ The curve illustrates the **trade-off** between the ratios of false positives and false negatives

3. Performance evaluation

Example of ROC curve



3. Performance evaluation

Choosing the best classifier

- **Holdout cross-validation**: the set of examples is randomly split k times into training, validation and test sets. The classifier that yields the best results over the k validation sets is chosen. We report its performance over the k test sets.
- **k -fold cross-validation**: we split the data into k equally sized subsets. Then k learning rounds are carried out, so that each subset serves once as validation set and once as test set. Again we choose the best classifier according to the validation performance, and we report the test performance.
- **Leave-one-out cross-validation**: an extreme form of k -fold cross-validation with $k=N$



4. NAÏVE BAYES CLASSIFIERS

4. Naïve Bayes classifiers

Definition

- Naïve Bayes classifiers are a kind of probabilistic classifiers which are based on Bayes' theorem
 - They assume that the input components x_i are independent given the class y

$$P(y | \mathbf{x}) = \frac{P(y)P(\mathbf{x} | y)}{\sum_{v \in V} P(v)P(\mathbf{x} | v)} \quad P(\mathbf{x} | y) = \prod_{d=1}^D P(x_d | y)$$

4. Naïve Bayes classifiers

Probability estimation

- In order to obtain an estimate of the **posterior probabilities** $P(x_d | y)$ the **m-estimate** can be used, where n' is the number of training examples of class y with x_d , n is the number of training examples of class y , p is a prior probability estimate (we might assume a uniform distribution), and m is a constant which expresses our confidence in p , measured in number of samples

$$P(x_d | y) \approx \frac{n' + mp}{n + m}$$

4. Naïve Bayes classifiers

ROC curve

- For a binary classification problem, we may use a variable **probability threshold** τ to assign an example to the positive class. This produces a ROC curve
 - The higher τ , the higher the true positive rate and the false positive rate

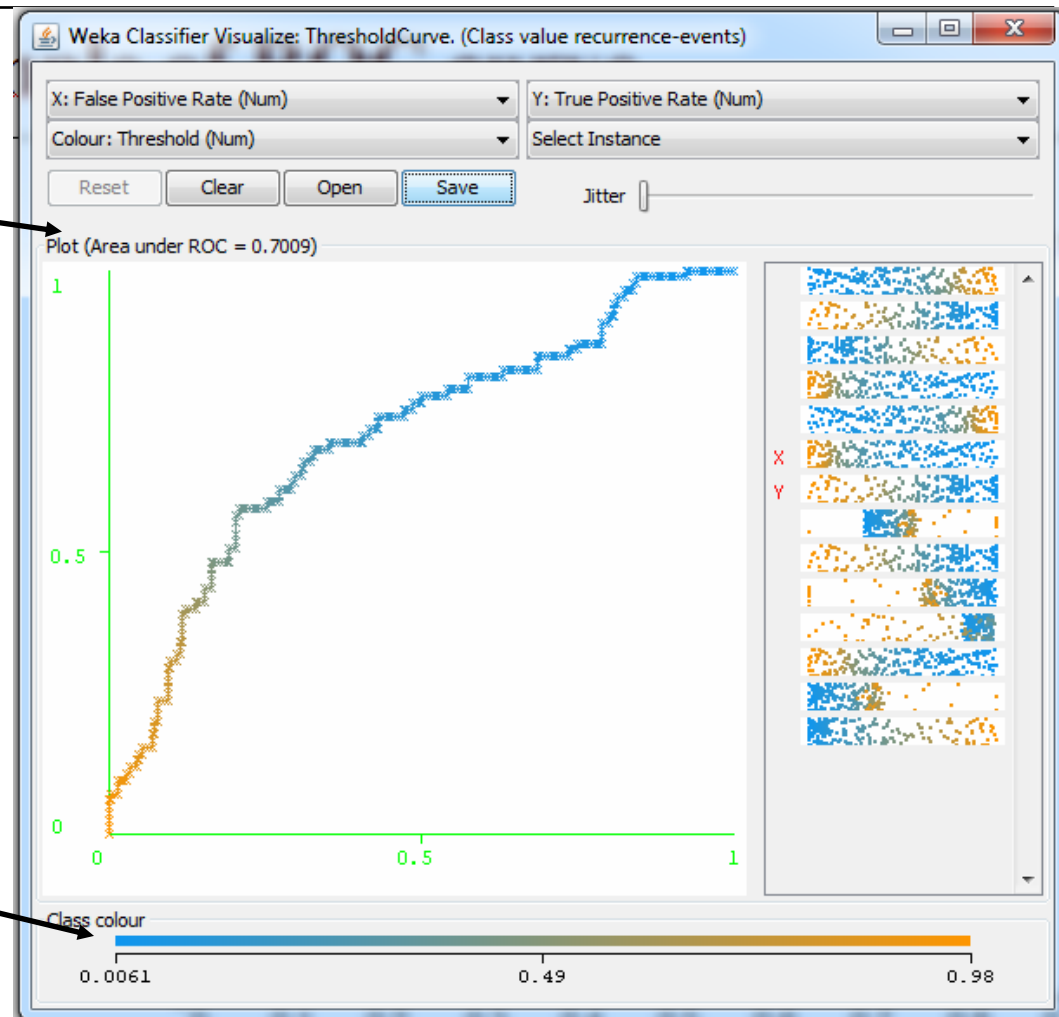
$$y_q = 1 \Leftrightarrow P(y = 1 | \mathbf{x}_q) > \tau$$

4. Naïve Bayes classifiers

ROC curve for breast-cancer dataset


Area under
ROC curve

Threshold τ





5. CONCLUSION



5. Conclusion

Summary

- ❑ **Classification** is a particular **supervised learning** task
- ❑ **Performance measures** are available for binary and multiclass problems
- ❑ **Cross-validation** techniques allow to choose a classifier and report its performance
- ❑ The **naïve Bayes classifiers** are both simple and powerful



5. Conclusion

Aftermath

- ❑ MYCIN was one of the first expert systems in history (Stanford University, 1970s)
 - It identified bacteria causing severe infections, and recommended antibiotics
- ❑ Nowadays Clinical Decision Support Systems (CDSS) are used all over the world as a valuable second opinion
 - It has been found that the use of these systems leads to significant improvements with respect to human-only procedures